A Bayesian hierarchical model for analysis of SNP diversity in multilocus, multipopulation samples

Feng Guo, Dipak K. Dey, and Kent E. Holsinger

June 28, 2008

Author's Footnote:

Feng Guo is Assistant Professor of Statistics, Department of Statistics, Virginia Tech, Blacksburg, VA 24061 (email: feng.guo@vt.edu), Dipak K. Dey is Professor and Head, Department of Statistics (email: dipak.dey@uconn.edu), and Kent E. Holsinger is Professor of Biology, Department of Ecology and Evolutionary Biology (email: kent@darwin.eeb.uconn.edu), University of Connecticut, Storrs, CT 06269. This research was supported, in part, by a grant from the National Institutes of Health, National Institute of General Medical Sciences 1R01-GM068449-01A1.

Abstract

The distribution of genetic variation among populations is conveniently measured by Wright's F_{ST}, which is a scaled variance taking on values in [0,1]. For certain types of genetic markers, and for single-nucleotide polymorphisms (SNPs) in particular, it is reasonable to presume that allelic differences at most loci are selectively neutral. For such loci, the distribution of genetic variation among populations is determined by the size of local populations, the pattern and rate of migration among those populations, and the rate of mutation. Because the demographic parameters (population sizes and migration rates) are common across all autosomal loci, locus-specific estimates of F_{ST} will depart from a common distribution only for loci with unusually high or low rates of mutation or for loci that are closely associated with genomic regions having a relationship with fitness. Thus, loci that are statistical outliers showing significantly more among-population differentiation than others may mark genomic regions subject to diversifying selection among the sample populations. Similarly, statistical outliers showing significantly less differentiation among populations than others may mark genomic regions subject to stabilizing selection across the sample populations. We propose several Bayesian hierarchical models to estimate locus-specific effects on F_{ST}, and we apply these models to single nucleotide polymorphism data from the HapMap project. Because loci that are physically associated with one another are likely to show similar patterns of variation, we introduce conditional autoregressive models to incorporate the local correlation among loci for high-resolution genomic data. We estimate the posterior distributions of model parameters using Markov chain Monte Carlo (MCMC) simulations. Model comparison using several criteria, including DIC and LPML, reveals that a model with locus- and population-specific effects is superior to other models for the data used in the analysis. To detect statistical outliers we propose an approach that measures divergence between the posterior distributions of locus-specific effects and the common F_{ST} with the Kullback-Leibler divergence measure. We calibrate this measure by comparing values with those produced from the divergence between a biased and a fair

coin. We conduct a simulation study to illustrate the performance of our approach for detecting loci subject to stabilizing/divergent selection, and we apply the proposed models to low- and high-resolution SNP data from the HapMap project. Model comparison using DIC and LPML reveals that CAR models are superior to alternative models for the high resolution data. For both low and high resolution data, we identify statistical outliers that are associated with known genes.

KEY WORDS: Bayesian approach, Hierarchical model, SNP, Wright's F_{st}, MCMC

1. INTRODUCTION

Human genetic diversity reflects our common evolutionary history. Differences among individuals belonging to the same group are smaller than those of individuals belonging to different groups. Moreover, differences among groups derived from the same broad geographical region are smaller than those derived from different geographical regions. For example, an analysis of microsatellite variation at 377 loci in 52 human populations (Rosenberg, Pritchard, Weber, Cann, Kidd, Zhivotovsky and Feldman 2002) identified five broad geographical clusters of populations: Africa, Eurasia, southeast Asia, Oceania, and the Americas. Approximately 75% of the among-population variation in allele frequency is associated with differences among these major geographical regions (Song, Dey and Holsinger 2006).

For loci that do not affect survival or reproduction, i.e., loci that are selectively neutral, both the amount of variation within populations and the extent of differentiation among populations are determined by: (1) the number of individuals in local populations, (2) the rates of migration among local populations, and (3) the mutation rates among alleles (see, for example, Crow and Kimura (1970), Fu, Gelfand and Holsinger (2003); Song et al. (2006)). In a typical population sample, individuals are genotyped at many loci, and all individuals are genotyped for the same set of loci. Thus, whatever the vagaries of demographic history – including population decline or expansion, population bottlenecks, asymmetric or variable migration rates among populations, etc. – all autosomal loci in a sample will share

that history, and they should show similar patterns of within- and among-population variation. Nonetheless, previous surveys of single-nucleotide polymorphisms (SNPs) in the human genome have revealed substantial differences among loci in the amount of among-population variation (see, for example, Akey, Zhang, Zhang, Jin and Shriver (2002) and Weir, Cardon, Anderson, Nielsen and Hill (2005)). Such differences suggest either that the mutational process differs substantially from locus to locus or that allelic differences at those loci (or loci with which they are closely associated) contribute differently to survival and reproduction than do allelic differences at other loci. Mutation rates at different SNP loci within a sample are likely to be comparable (Chakraborty, Kimmel, Stivers, Davison and Deka 1997; Weber and Wong 1993) (but see Lercher and Hurst (2002) for a cautionary note). Thus, if a few SNP loci show substantially more among-population differentiation than the rest, these loci may mark regions of the genome at which there has been divergent selection across populations in the sample. Similarly, SNP loci showing substantially less differentiation may mark regions subject to stabilizing selection across populations.

Cavalli-Sforza (1966) may have been the first to suggest using measures of population divergence to detect natural selection, but Lewontin and Krakauer (1973) were the first to propose using Wright's F-statistics for this purpose. Nei and Maruyama (1975) and Robertson (1975) quickly pointed out that comparing a point estimate of F_{ST} for a particular locus with a point estimate for the genomic background fails to account for the large variance in F_{ST} among loci expected as a result of genetic drift, variance that is intrinsic to the stochastic evolutionary process and that cannot be eliminated by increased sampling. Nonetheless, Beaumont and Balding (2004) showed that Bayesian p-values derived from locus-specific F_{ST} estimates could be used to detect statistical outliers that corresponded to loci under selection in their simulations. Recently, Riebler, Held and Stephan (2008) extended this approach by introducing binary indicator variables whose posterior can be used to identify statistical outliers.

We take a similar approach. To identify loci that show unusually large or unusually small amounts of differentiation at SNP loci, we develop hierarchical Bayesian models for analysis

of multilocus, multipopulation SNP data, and we combine them with a novel approach to identify loci that are statistical outliers. Hierarchical models are natural in this context because the underlying patterns of biological diversity are hierarchical. In this paper, populations are predefined based on the geographic origin of samples. Conceptually, we assume that populations have diverged from a common ancestor, a hyperpopulation. Consequently, we assume the allele frequency at each SNP locus in each population is drawn from a common hyperpopulation in which allele frequencies vary across loci. Although all autosomal loci have the same expected value of F_{ST}, the population sample at each locus represents a different realization from a stochastic evolutionary process and the realized F_{ST} at each locus will be different. Thus, we assume that F_{ST} at any particular locus is drawn from a hyperdistribution. The variability of this hyperdistribution reflects the among-realization variability in the stochastic evolutionary process. Loci with substantially greater or substantially smaller amounts of among-population differentiation than are consistent with this hyperdistribution will be identified as outliers. Thus, our inference is based on comparing the posterior distribution of a parameter reflecting locus-specific effects on Wright's F_{ST} (namely, $\theta_i, i = 1, \ldots, I$) with the posterior distribution of parameters reflecting a genome-wide distribution for F_{ST}.

Specifically, we characterize both the common posterior distribution that describes most loci in the sample and the posterior distribution of each θ_i. In the hierarchical structure we propose, the hyperprior for θ_i is given by a beta distribution with mean φ and variance $\varphi(1-\varphi)\theta_L$. Thus, a beta distribution with mean $\hat{\varphi}$ and variance $\hat{\varphi}(1-\hat{\varphi})\hat{\theta}_L$ is a suitable choice for the common posterior distribution, where $\hat{\varphi}$ and $\hat{\theta}_L$ refer to posterior means. The posterior distributions of the θ_i are unimodal and have support on $[0, 1]$. Thus, we approximate them with a beta distribution by matching the posterior means and variances. By using this approach, we have a closed form for the posterior density function, and we can use the well accepted Kullback-Leibler divergence (KLD) measure to compare the posterior distribution of each θ_i with the common posterior distribution specified by $\hat{\varphi}$ and $\hat{\theta}_L$. We calibrate this divergence using the method proposed by Peng and Dey (1995).

There are 23 chromosomes in the human genome and approximately 3.2M SNP loci in the dataset from which our data set is derived. Adjacent loci in the complete data set are separated by an average of only 1000 nucleotides. Thus, some loci are in close physical proximity, and it is reasonable to expect that they will show similar patterns of variation as a result. We introduce a conditional autoregressive (CAR) model for high-resolution genomic data to incorporate the effects of physical proximity among the loci into the model. The proximity effects are brought into the model by constructing a CAR prior for θ_i. Thus, we consider four models in this paper: (1) a hierarchical model in which we assume that the θ_i are random samples from a hyperdistribution for F_{ST}, (2) a hierarchical model in which we use a product decomposition to distinguish locus- and population-specific effects on F_{ST}, (3) a CAR extension of model 1, and (4) a CAR extension of model 2. Because there are no closed form expressions for the posterior distribution of θ_i, we use a sampling based Markov chain Monte Carlo (MCMC) method to obtain the marginal posterior distributions.

The remainder of the paper is structured as follows: In section 2 we introduce the genetic data used in the analysis; section 3 provides a detailed description of the models; the method of detecting loci that are statistical outliers is described in section 4; in section 5, a simulation study is conducted to demonstrate the proposed models; the application and model comparison are presented in section 6; and section 7 provides a summary of the main results and discusses their implications.

2. THE SNP DATA

Whether we can identify certain loci as having unusually large or unusually small amounts of among-population differentiation depends on both the number of populations included in the sample and on the number of loci scored per individual. Because F_{ST} is directly proportional to the allele frequency variance among populations, the precision of F_{ST} estimates and our ability to detect outliers will obviously increase as the number of populations included in the sample increases. Moreover, the greater the number of populations included in the sample, the greater the chances that one or more of them have been subject to divergent

selection leading to divergent allele frequencies. Similarly, the larger the number of markers included in a data set, the more precisely we are able to estimate the amount of among-locus variability in F_{ST} that is expected and the more power we have to detect loci that depart significantly from the common distribution.

We analyze publicly available data from the HapMap project (Consortium 2005), which provides data on *circa* 3.2M polymorphic SNPs. These data are derived from a relatively small number of individuals (270) and only four populations: Yoruba in Ibadan, Nigeria (YRI); Japanese in Tokyo, Japan (JPT); Han Chinese in Beijing, China (HCB); and Utah residents with ancestry from northern and western Europe (CEPH). Thus, the HapMap data provide an opportunity for high-resolution analysis of variation patterns across the human genome, but the small number of populations included in the sample will allow us to identify only those loci in which departures from the common distribution are especially large.

For the notation we assume individuals are sampled from K populations. By "population" we refer to sampling location. Both the Yoruba population sample (YRI) and the U.S. population sample (CEPH) consist of 30 trios: two parents and one offspring. To avoid modeling the dependence structure this sampling would induce, we analyze only parental genotypes in YRI and CEPH. For each individual, the genotype is determined at I SNP loci. Because nearly all SNP loci have only two alleles, we label alleles as A_1 and A_2 at each locus. As will become evident in the model description, inference on F_{ST} does not depend on the labeling of alleles. The data are aggregated to allele counts by locus and population. Denote x_{ik} as the sample size of allele A_1 and N_{ik} as the total number of alleles sampled at locus i in population k. Obviously, the sample size of allele A_2 at locus i for population k is $N_{ik} - x_{ik}$.

In order to implement the CAR model, we also need a proximity or adjacency matrix, $\mathbf{W}$, in which element w_{ij} represents the spatial dependence between locus i and j. We consider distances measured in terms of the frequency of recombination between the markers. To calculate recombinational distances we used map positions (measured in centimorgans) as estimated by Peter Donnelly, Gil McVean, and Simon Myers in a dataset available for

download from the HapMap site.

We focus our attention on SNP loci on human chromosome 7, for which 201,656 SNP loci were scored in the four populations included in the HapMap data set. The number of alleles in each population varies from around 70 to 120. Some populations are completely lacking genotypes at particular SNP loci. We included only those loci for which genotypes counts were available for all populations. The pruned data set includes $177,374$ loci. We also exclude a small number of loci in which all populations are fixed for one allele, i.e., loci for which the frequency of one allele is zero in all populations. At such loci there is no among-population variation in allele frequency to assess. Loci that are monomorphic in all populations may mark genomic regions subject to strong stabilizing selection. Thus, by excluding these loci from our analysis we reduce our ability to detect loci showing unusually small amounts of divergence among populations. To screen the whole chromosome while keeping the computational demands reasonable, we first perform a low-resolution scan by selecting loci throughout chromosome 7 but including only every 50th locus. There are 3040 loci separated by 52,000 nucleotides on average included in the final analysis. We then focus on a region marked by a strong statistical outlier in the low-resolution scan and perform a high-resolution scan that includes 3002 loci at intervals of approximately 860 nucleotides.

3. MODELS

3.1 Describing genetic structure

First consider one locus with multiple alleles. Let $p_{m,k}$ be the frequency of allele A_m ($m = 1,\ldots,M$) in population k ($k = 1,\ldots,K$), and assume that alleles are associated randomly within individuals (i.e., genotypes are in Hardy-Weinberg proportions) so that the frequency of the ordered genotype ($m \leq n$) A_mA_n in population k is given by

$$\gamma_{mn,k} = \begin{cases} p_{m,k}^2 & \text{for } m = n \\ 2p_{m,k}p_{n,k} & \text{for } m \neq n \end{cases} .$$

Then the mean genotype frequency, $\gamma_{mn\cdot}$, across the set of K populations is given by

$$\begin{aligned} \gamma_{mn\cdot} &= \frac{1}{K}\sum_{k=1}^{K}\gamma_{mn,k} \\ &= \begin{cases} p_{m\cdot}^2 + s_{p_m}^2 & \text{for } m = n \\ 2p_{m\cdot}p_{n\cdot} - 2s_{p_m p_n} & \text{for } m \neq n \quad , \end{cases} \end{aligned} \tag{1}$$

where $p_{m\cdot} = \frac{1}{K}\sum_{k=1}^{K} p_{m,k}$, $s_{p_m}^2 = \frac{1}{K}\sum_{k=1}^{K}(p_{m,k} - p_{m\cdot})^2$, and $s_{p_m p_n} = \frac{1}{K}\sum_{k=1}^{K}(p_{m,k} - p_{m\cdot})(p_{n,k} - p_{n\cdot})$ (see Li (1955)). If alleles are exchangeable in the underlying stochastic evolutionary process, the allele frequencies are identically distributed at stationarity under quite general conditions (Fu et al. 2003). Specifically, $E(p_{m\cdot}) = \pi$, $E(s_{p_m}^2) = \sigma_p^2$, and $E(s_{p_m p_n}) = \rho\sigma_p^2$, where π, σ_p^2, and ρ are the common values (Fu et al. 2003). Under these conditions the expectation of the $\gamma_{mn\cdot}$ can be written as

$$E(\gamma_{mn\cdot}) = \begin{cases} \pi^2 + F_{st}\pi(1-\pi) & \text{for } m = n \quad , \\ 2\pi(1-\pi)(1-F_{st}) & \text{for } m \neq n \quad , \end{cases} \tag{2}$$

where

$$F_{st} = \frac{\sigma_p^2}{\pi(1-\pi)} \quad . \tag{3}$$

Since the work of Wright (1951) and Malécot (1948), F_{ST} has been the most widely used statistic for summarizing patterns of among-population differentiation in population genetics.

Assume that we have a sample of allelic data from I loci. For notational simplicity we restrict our attention to the case where each locus has only two alleles. The models discussed in this paper can be relatively easily extended to multiple alleles, and an outline of the extension is introduced in Holsinger (1999). Let x_{ik} denote the count of allele A_1 in the sample from locus i in population k, let N_{ik} be the total number of alleles sampled at locus i in population k, and let p_{ik} be the "true" allele frequency at locus i of population k. Then the first-stage likelihood is a product binomial:

$$f(\mathbf{x} \mid \mathbf{p}) \propto \prod_{i=1}^{I}\prod_{k=1}^{K} p_{ik}^{x_{ik}}(1-p_{ik})^{N_{ik}-x_{ik}} \quad . \tag{4}$$

We assume that the distribution of allele frequencies among populations at locus i is a beta distribution with parameters $((1-\theta_x)/\theta_x)\pi_i$ and $((1-\theta_x)/\theta_x)(1-\pi_i)$ and that the distribution of allele frequencies across loci is a beta distribution with parameters $((1-\theta_y)/\theta_y)\pi$ and $((1-\theta_y)/\theta_y)(1-\pi)$. It is straightforward to show that this hierarchical structure produces a mean and covariance structure that matches (2) and (3) (Holsinger 2006; Song et al. 2006). While the stationary distribution of among-population allele frequencies follows a beta distribution in some evolutionary models (Crow and Kimura 1970), we make no explicit assumption about the underlying evolutionary process in using a beta distribution to describe variation in allele frequencies among populations. We adopt it simply because it is a flexible distribution suitable for many distributions on [0,1]. Indeed, in a dataset with samples from a large number of populations it may be desirable to consider a finite mixture of beta distributions to allow for multimodality in the allele frequency distribution. Placing vague, uniform priors on π, θ_x, and θ_y completes the specification of a Bayesian model and allows us to construct an MCMC sampler for inference on the parameters. In particular, $\theta_x(1-\theta_y)+\theta_y$ is mathematically equivalent to F_{ST} as estimated in Weir and Cockerham (1984)'s random effect model. F_{ST} provides a convenient measure of genetic differentiation among populations, because it is interpretable as the proportion of genetic diversity due to allele frequency differences among populations. Different demographic histories, different local population sizes, and different patterns of migration will lead to different amounts of among-population differentiation and to correspondingly different values F_{ST}, but all autosomal loci within an individual will be affected in the same way. Thus, all autosomal loci in a typical population sample will have values of F_{ST} drawn from the same distribution unless rates of mutation or patterns of selection differ substantially. As in Akey et al. (2002), Beaumont and Balding (2004), Storz, Payseur and Nachman (2004), and Weir et al. (2005), we shall use locus-specific estimates of F_{ST} to detect loci showing divergent patterns of variation.

All 4 models proposed in this paper are based on the framework introduced above. Directed acyclic graphs (DAG) showing the structure of each model are shown in Figure 1.

[Figure 1 about here.]

3.2 Model 1

The first model proposed here extends the simple framework above by incorporating locus-specific estimates of F_{ST}. As discussed above, all loci have 2-alleles and the likelihood of the data is a product binomial distribution. We place a beta prior with parameters $((1-\theta_i)/\theta_i)\,\psi_i, (1-\theta_i)/\theta_i)\,(1-\psi_i)$ for the binomial parameter p_{ik}. It can be easily shown that the expectation of p_{ik} in the prior distribution is ψ_i and that its variance is $\theta_i\psi_i(1-\psi_i)$. Thus, θ_i corresponds directly with Wright's F_{ST} for locus i. We adopt a full Bayesian approach and set the second and third level hierarchical prior for θ_i and ψ_i respectively. The posterior distribution is as follows:

$$\pi(\Theta|\mathbf{D}) \propto f(\mathbf{x}|\boldsymbol{p})\pi(\boldsymbol{p}|\boldsymbol{\theta},\boldsymbol{\psi})\pi(\boldsymbol{\theta}|\theta_L,\varphi)\pi(\boldsymbol{\psi}|\psi_H,\nu)\pi(\theta_L)\pi(\varphi)\pi(\psi_H)\pi(\nu),$$

where Θ is the collection of all the model parameters; $f(\mathbf{x}|\boldsymbol{p})$ is the likelihood function as in (4). The first level prior for $\boldsymbol{p}$ is,

$$\pi(\boldsymbol{p}|\boldsymbol{\theta},\boldsymbol{\psi}) = \prod_{i=1}^{I}\prod_{k=1}^{K} \frac{\Gamma(\frac{1-\theta_i}{\theta_i})}{\Gamma(\frac{1-\theta_i}{\theta_i}\psi_i)\Gamma(\frac{1-\theta_i}{\theta_i}(1-\psi_i))} p_{ik}^{\frac{1-\theta_i}{\theta_i}\psi_i-1}(1-p_{ik})^{\frac{1-\theta_i}{\theta_i}(1-\psi_i)-1},$$

which is a product of beta distributions with parameters $(\frac{1-\theta_i}{\theta_i}\psi_i, \frac{1-\theta_i}{\theta_i}(1-\psi_i))$. The hyper-parameter θ_i corresponds to F_{ST} at locus i and is the key parameter of interest. We place a second level of prior on $\boldsymbol{\theta}$ as follows:

$$\pi(\boldsymbol{\theta}|\theta_L,\varphi) = \prod_{i=1}^{I} \frac{\Gamma(\frac{1-\theta_L}{\theta_L})}{\Gamma(\frac{1-\theta_L}{\theta_L}\varphi)\Gamma(\frac{1-\theta_L}{\theta_L}(1-\varphi))} \theta_i^{\frac{1-\theta_L}{\theta_L}\varphi-1}(1-\theta_i)^{\frac{1-\theta_L}{\theta_L}(1-\varphi)-1}, \tag{5}$$

which is a beta prior with parameters $(\frac{1-\theta_L}{\theta_L}\varphi, \frac{1-\theta_L}{\theta_L}(1-\varphi))$.

The second level prior for $\boldsymbol{\psi}$ is a beta distribution with parameters $(\frac{1-\nu}{\nu}\psi_H, \frac{1-\nu}{\nu}(1-\psi_H))$,

$$\pi(\boldsymbol{\psi}|\nu,\psi_H) = \prod_{i=1}^{I} \frac{\Gamma(\frac{1-\nu}{\nu})}{\Gamma(\frac{1-\nu}{\nu}\psi_H)\Gamma(\frac{1-\nu}{\nu}(1-\psi_H))} \psi_i^{\frac{1-\nu}{\nu}\psi_H-1}(1-\psi_i)^{\frac{1-\nu}{\nu}(1-\psi_H)-1}. \tag{6}$$

At the third level, there is no preference for any particular value. We use a Uniform(0,1) prior for θ_L, ν, ψ_H, and φ.

3.3 Model 2

The second model proposed is an extension of the first model. We replace θ_i in Model 1 with θ_{ik} for locus i and population k, where $\theta_{ik} = 1 - (1 - \theta_i)(1 - \theta_k)$. In this formulation θ_i represents a locus-specific effect and θ_k represents a population-specific effect. As in Model 1, hierarchical beta priors are assigned to θ_i and θ_k respectively. The posterior distribution is as follows:

$$\pi(\Theta|\mathbf{D}) \propto f(\mathbf{x}|\boldsymbol{p})\pi(\boldsymbol{p}|\boldsymbol{\theta}, \boldsymbol{\theta_k}, \boldsymbol{\psi})\pi(\boldsymbol{\theta}|\theta_L, \varphi)\pi(\boldsymbol{\theta_k}|\theta_P, \phi)\pi(\boldsymbol{\psi}|\psi_H, \nu)\pi(\theta_L)\pi(\varphi)\pi(\theta_P)\pi(\phi)\pi(\psi_H)\pi(\nu)$$

where $\boldsymbol{\theta_k}$ is the vector of $\theta_k, k = 1, \ldots, K$. We further assume parameters $\boldsymbol{\theta_k}$ come from a hyper-beta distribution with the following form:

$$\pi(\boldsymbol{\theta_k}|\theta_P, \phi) = \prod_{k=1}^{K} \frac{\Gamma(\frac{1-\theta_P}{\theta_P})}{\Gamma(\frac{1-\theta_P}{\theta_P}\phi)\Gamma(\frac{1-\theta_P}{\theta_P}(1-\phi))} \theta_k^{\frac{1-\theta_P}{\theta_P}\phi-1}(1-\theta_k)^{\frac{1-\theta_P}{\theta_P}(1-\phi)-1}.$$

The prior for $\boldsymbol{\theta}$, $\boldsymbol{\psi}$ is the same as in model 1, equations (5) and (6). Further, priors for θ_L, θ_P, ν, φ, ϕ, and ψ_H are assumed Uniform(0,1).

3.4 CAR Model

The hierarchical models discussed above allow the θ_i to "borrow strength" from other sites in estimating their posterior distributions, but they treat variation at each locus as if it were independent of variation at all other loci. Analysis of locus-specific effects at high genomic resolution is essential if the results of our method are to provide experimentalists with a guide for selecting regions worthy of additional study. But at high resolutions the reduced probability of recombination among adjacent SNP markers is likely to lead to similar patterns of differentiation among populations, i.e., we expect the locus-specific effects of neighboring loci on F_{ST} to be similar. To account for this correlation, we adopt a common methodology used in the analysis of spatial variation in geographical models, namely a conditional autoregressive (CAR) model on random effects associated with each locus. The basic idea is that the loci close to each other are more likely to have similar amounts of among-population differentiation and thus similar posterior distributions for θ_i. Specifically,

we incorporate the local correlation into the model through a CAR prior for θ_i in Model 1 and Model 2.

We incorporate the local correlation structure into the hierarchical model 1 and 2 by placing a prior distribution with CAR components on θ_i, and we use a logit transformation to extend the support of θ_i to entire real line. In short, the model specification is as follows:

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \mu + \epsilon_i, i = 1, 2, \ldots I. \tag{7}$$

Here μ captures the global mean and ϵ_i represents a random effect associated with locus i. We place a normal prior with mean zero and variance $1/\tau_h$ on μ, i.e.,

$$\mu \sim N(0, 1/\tau_h). \tag{8}$$

We place a CAR prior on ϵ_i to incorporate the local correlation among loci. The CAR prior reflects our expectation that at high genomic resolution ϵ_i and $\epsilon_{i'}$ will be of similar sign and magnitude when the genetic distance between i and i' is small. The CAR prior has the following conditional form:

$$\epsilon_i | \boldsymbol{\epsilon}_{(-i)} \sim N\left(\sum_{j \neq i} \frac{w_{ij}}{w_{i+}} \epsilon_j, \frac{1}{\tau_c w_{i+}}\right), i = 1, 2, \ldots I, \tag{9}$$

where $\boldsymbol{\epsilon}_{(-i)}$ is the collection of $\epsilon_j, \forall j \neq i$, τ_c is a precision parameter, w_{ij} is the entry at row i and column j of proximity matrix $\mathbf{W}$, and $w_{i+} = \sum_{j=1}^{I} w_{ij}$. The $\mathbf{W}$ is an $I \times I$ proximity matrix in which entry w_{ij} indicates the spatial relationship between loci i and j. Several choices of $\mathbf{W}$ are possible. A simple choice would be to use 0 or 1 to indicate whether or two loci are "close" or not to each other, where "close" is defined as being within a certain distance. Because we expect the statistical association among loci to be related to the recombinational distance between loci, we define w_{ij} as a function of the distance between loci,

$$w_{ij} = \begin{cases} c(d_{ij}) & \text{if loci } i \neq j \\ 0 & \text{if loci } i = j, \end{cases}$$

where d_{ij} is the distance between loci i and j, and $c(d_{ij})$ is a function that describes how the covariance among loci depends on the distance between them. The $c(d_{ij})$ is usually a decreasing function of d_{ij}, often a reciprocal or an exponential. Because the recombinational distance between some of our loci in the high-resolution scan is zero, we use the exponential function, $c(d_{ij}) = c_1 + c_2 \exp(-c_3 d_{ij})$, where c_1, c_2, c_3 are constants chosen for computational convenience and numerical stability. We chose the value of c_1, c_2, c_3 so that (1) only the 20-100 nearest loci have a large influence; (2) the average value of w_{i+} is approximately 1; and (3) w_{i+} is greater then 0.5 to avoid numerical instability associated with small values of w_{i+}.

The joint prior distribution for the ϵ_i is

$$\pi(\epsilon_1, \ldots, \epsilon_I) \propto \exp\left\{ -\frac{\tau_c}{2} \sum_{i \neq j} \omega_{ij} (\epsilon_i - \epsilon_j)^2 \right\}.$$

Note that this is a pairwise difference model and is not a proper distribution (Banerjee, Carlin and Gelfand 2004). In particular, the ϵ_i are nonidentifiable. As usual in such models, the constraint $\sum \epsilon_i = 0$ is sufficient to guarantee identifiability. Here θ_i can be calculated from the inverse logit function

$$\theta_i = \frac{e^{\mu+\epsilon_i}}{1 + e^{\mu+\epsilon_i}}, \text{ and } \frac{1-\theta_i}{\theta_i} = e^{-(\mu+\epsilon_i)}.$$

By replacing the θ_i in Model 1 with the μ and ϵ_i, the prior for p_{ik} is then reduced to,

$$\begin{aligned} &\pi(p_{ik}|\epsilon_i, \mu, \psi_i) \\ &= \frac{1}{B(e^{-(\mu+\epsilon_i)}\psi_i, e^{-(\mu+\epsilon_i)}(1-\psi_i))} p_{ik}^{e^{-(\mu+\epsilon_i)}\psi_i - 1} (1-p_{ik})^{e^{-(\mu+\epsilon_i)}(1-\psi_i)-1}, \end{aligned}$$

where $B(\cdot, \cdot)$ denotes the beta function. The rest of the model specification is the same as Model 1.

The last model proposed uses the CAR prior for θ_i in Model 2. Again, we use a logit transformation for θ_i and a random effect model as in (7), (8), and (9). Then using the

identity $(1-\theta_{ik}) = (1-\theta_i)(1-\theta_k)$ we have,

$$\theta_{ik} = 1 - \frac{1-\theta_k}{1+e^{u+\epsilon_i}}, \text{ and } \frac{1-\theta_{ik}}{\theta_{ik}} = \frac{1-\theta_k}{e^{\mu+\epsilon_k}+\theta_k}.$$

Thus the prior for p_{ik} is obtained as,

$$\pi(p_{ik}|\mu\,\epsilon_i, \theta_k, \psi_i) = \frac{1}{B\left(\frac{1-\theta_k}{\exp(\mu+\epsilon_i)+\theta_k}\psi_i, \frac{1-\theta_k}{\exp(\mu+\epsilon_i)+\theta_k}(1-\psi_i)\right)} p_{ik}^{\frac{1-\theta_k}{\exp(\mu+\epsilon_i)+\theta_k}\psi_i - 1}(1-p_{ik})^{\frac{1-\theta_k}{\exp(\mu+\epsilon_i)+\theta_k}(1-\psi_i)-1}.$$

The rest of the model specification is the same as in Model 2.

As recommended by Banerjee et al. (2004), we adopt the following prior distributions for τ_c and τ_h: $\tau_h \sim \text{Gamma}(0.001, 0.001)$, and $\tau_c \sim \text{Gamma}(0.1, 0.1)$.

4. DETECTING LOCI WITH UNUSUAL PATTERNS OF VARIATION

The overarching objective of our models is to allow us to identify loci that are "unusual," i.e., loci for which the amount of among-population variation differs substantially from that at other loci. Statistically, this corresponds to identifying loci for which θ_i is either unusually large or unusually small. As Nei and Maruyama (1975) and Robertson (1975) pointed out more than 30 years ago, however, it is not sufficient to ask whether a particular θ_i is different from a common mean. Such a comparison would account only for the statistical uncertainty associated with parameter estimates. It would neglect the much larger uncertainty often associated with the underlying stochastic evolutionary process. In our approach, we assume that the θ_i are drawn independently from a common hyperdistribution. Thus, if all loci in the sample were selectively neutral, the variability among loci in θ_i captured by this hyperdistribution would reflect variability in outcomes associated with different realizations of the underlying stochastic evolutionary process. If mutation rates differ among loci, that variation will also be reflected in the variability of this hyperdistribution. Thus, to detect θ_i that are unusually large or unusually small, we must compare them with a common distribution rather than a common mean.

We propose the following steps to detect outliers: (1) Approximate the posterior distribution of locus-specific effect parameters, i.e., θ_i for the beta-hierarchical model and ϵ_i for the CAR models (see next paragraph for details). (2) Calculate the distance between the locus-specific effect and a "centering" distribution derived from the hyperdistribution describing among-locus variation in the locus-specific effect. (3) Compare the mean of the posterior distribution for loci identified as outliers with the mean of the "centering" distribution to identify loci with unusually large or unusually small amounts of among-population differentiation.

Our preliminary analysis shows that the posterior distribution of θ_i is unimodal. It is well known that any unimodal distribution with support on $[0, 1]$ can be approximated by a beta distribution. Therefore, we approximate the posterior distribution of θ_i with a beta distribution whose first two moments match the first two moments of the posterior distribution for θ_i as estimated from the MCMC output. We compare the posterior distribution for each θ_i with the posterior of its hyperdistribution. For example, in Model 1, each θ_i is compared with

$$\text{Beta}\left(\frac{1-\hat{\theta_L}}{\hat{\theta_L}}\hat{\varphi}, \frac{1-\hat{\theta_L}}{\hat{\theta_L}}(1-\hat{\varphi})\right) \quad ,$$

where $\hat{\varphi}$ and $\hat{\theta}_L$ refer to posterior means. The loci for which the posterior of θ_i diverges substantially from this hyper-distribution are considered as outliers.

We use the KLD to measure the divergence between the posterior of θ_i and its centering distribution. The KLD between two densities $p(y)$ and $q(y)$ is defined as

$$\text{KLD}(p, q) = \int p(y) \log\left(\frac{p(y)}{q(y)}\right) dy \quad .$$

The KLD between two beta distributions with parameters (α_0, β_0) and (α_1, β_1) is given by

$$\text{KLD} = \int \frac{1}{B(\alpha_0, \beta_0)} \theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1} \log \frac{\frac{1}{B(\alpha_0, \beta_0)} \theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1}}{\frac{1}{B(\alpha_1, \beta_1)} \theta^{\alpha_1 - 1}(1-\theta)^{\beta_1 - 1}} d\theta.$$

If $X \sim \text{Beta}(\alpha, \beta)$ then $1-X \sim \text{Beta}(\beta, \alpha)$. Furthermore, if $X \sim \text{Beta}(\alpha, \beta)$, then $E[\log X] = \psi(\alpha) - \psi(\alpha + \beta)$, where $\psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ is the digamma function. Thus, $E[\log(1 - X)] =$

$\psi(\alpha+\beta)-\psi(\beta)$, and the KLD between two beta distribution is

$$\text{KLD} = \log\frac{B(\alpha_1,\beta_1)}{B(\alpha_0,\beta_0)} + (\alpha_0-\alpha_1)(\psi(\alpha_0)-\psi(\alpha_0+\beta_0)) + (\beta_0-\beta_1)(\psi(\beta_0)-\psi(\alpha_0+\beta_0)),$$

where $\psi(\cdot)$ is the digamma function.

Similarly, we compare the posterior distribution of each locus-specific effect from the CAR models, ϵ_i, with the posterior of its corresponding hyperdistribution. The "centering" distribution can be calculated in two ways, corresponding to detecting loci with unusually large or unusually small amounts of differentiation either relative to near neighbors ("local outliers") or relative to all loci in the sample ("global outliers"). The "centering" distribution for detecting local outliers is

$$\epsilon_i|\hat{\boldsymbol{\epsilon}}_{(-i)} \sim N\left(\sum_j \frac{w_{ij}}{w_{i+}}\hat{\epsilon}_j, \frac{1}{\hat{\tau}_c w_{i+}}\right), i = 1,2,\ldots I \quad , \tag{10}$$

where $\hat{\epsilon}_j$ and $\hat{\tau}_c$ refer to posterior means. Accordingly we define

$$\text{KLD}_{local} = \text{KLD}(\epsilon_i|\text{D}, \epsilon_i|\hat{\boldsymbol{\epsilon}}_{(-i)}), \tag{11}$$

where $\epsilon_i|\text{D}$ is the marginal posterior distribution of ϵ_i. We approximate this distribution with a normal distribution by matching the first two moments. So KLD_{local} provides a measure of the divergence between the posterior of ϵ_i and a locally smoothed estimate. A large KLD_{local} indicates a locus differing substantially from its near neighbors.

Recall that for identifiability of the model we impose the constraint $\sum_i \epsilon_i = 0$. Thus, a locus for which ϵ_i is substantially different from zero exhibits either substantially more or substantially less differentiation among populations than the average locus in the sample. In short, it also makes sense to compare $\epsilon_i|\text{D}$ with the marginal distribution, $N\left(0, \frac{1}{\hat{\tau}_c w_{i+}}\right)$. Accordingly we define

$$\text{KLD}_{global} = \text{KLD}\left(\epsilon_i|\text{D}, N\left(0, \frac{1}{\hat{\tau}_c w_{i+}}\right)\right) \quad . \tag{12}$$

KLD_{global} measures the divergence between the posterior distribution of θ_i and the mean among-population differentiation. A group of loci with large global KLD but small local

KLD indicates a cluster of loci with substantially more or substantially less among-population differentiation than the average locus in the sample. It is straightforward to show that the KLD between two normal distributions is

$$\mathrm{KLD}(N(\mu_0, \sigma_0^2), N(\mu_1, \sigma_1^2)) = \frac{1}{2}\left[\log\frac{\sigma_1^2}{\sigma_0^2} + \frac{\sigma_0^2}{\sigma_1^2} + \frac{1}{\sigma_1^2}(\mu_0 - \mu_1)^2 - 1\right].$$

We calibrate the KLD using the method proposed by Peng and Dey (1995). Consider flipping a "fair" coin with equal probability 0.5 for head and tail versus flipping a biased coin with probability θ for head. The larger $|\theta - 0.5|$ is, the more "extreme" the bias. The KLD measure between these two Bernoulli distributions is

$$L = \log(0.5) - 0.5 * \log(\theta(1-\theta)).$$

For example, $\theta = 0.01$, corresponds to a strong bias and a KLD value of 1.614.

The KLD value provides a measure of the distance between two distributions but no information about the relative locations of the centers of the two distributions. For example, two normal distributions, $N(-1, 3)$ and $N(1, 3)$, both have the same KLD relative to a standard normal distribution, $N(0, 1)$. Outlier detection thus follows a two-stage procedure. First, we identify loci with a large KLD between the posterior of the locus-specific effect, θ_i or ϵ_i and the corresponding centering distribution. Second, we compare the posterior means of θ_i or ϵ_i for loci identified as outliers and the means of the centering distribution to determine whether the locus shows unusually large or unusually small amounts of among-population differentiation.

5. SIMULATION STUDY

To determine whether outliers detected with our method correspond to loci subject to selection, we simulate allele frequencies under a Wright-Fisher model with migration, mutation, and drift, following Beaumont and Balding (2004). A small number of loci included in the simulation are also subject to natural selection. Specifically, we simulate a sample of allele frequencies drawn from four populations as in the SNP data from the HapMap project. We

assume a constant population size of 250 individuals (500 chromosomes) for all populations, and we assume that all sampled loci are independently inherited. The migration rate into a population is chosen by sampling F_{ST} from a beta distribution with parameters (0.25, 2.25) and setting $m = (1 - F_{ST})/(2NF_{ST})$, where N is the population size (see Beaumont and Balding (2004) for details). The chosen parameters result in a distribution of F_{ST} comparable to that observed in the HapMap data.

The simulation allows for three types of loci: those subject to directional (divergent) selection, those subject to balancing (stabilizing) selection, and neutral loci. Allelic differences at neutral loci do not affect fitness. Levels of within- and among-population variation are determined entirely by migration, mutation, and genetic drift. We assume that the majority of loci are selectively neutral in our simulations. Thus, variation at these loci largely determines the distribution of F_{ST} across loci.

At a locus under directional selection, one allele enhances the fitness of individuals carrying it. When the allele enhancing fitness differs among populations, allele frequency differences among populations will be greater than at neutral loci. At a locus under balancing selection, heterozygous individuals are more likely to reproduce than individuals homozygous for either allele. In our simulations, the loci are unlinked and each is either neutral, subject to divergent selection, or subject to balancing selection. In the case of loci subject to directional selection, the relative fitness is $1 + s$ for the favored homozygote, $1 + s/2$ for heterozygote, and 1 for the disfavored genotype. In the case of loci subject to balancing selection, the relative fitness of heterozygotes is $1 + s$ and the relative fitness of both homozygotes is 1.

We consider two different mutation models. In the two-locus model, the marker locus is completely linked to the locus that is under selection. The marker locus evolves according to an infinite allele model while the selected locus evolves according to a parent-independent K-allele model with three alleles. In the marker-selected model, the marker itself is subject to selection and evolves according to the parent-independent K-allele model with three alleles (see Beaumont and Balding (2004) for more details). The simulations were implemented

using software provided by Mark Beaumont. The mutation rate at marker loci is $\mu_m = 0.00001$ and at selected loci is $\mu_s = 0.0001$. We generated 100 samples from each population after $50,000$ generations in the simulation from 11 different simulations scenarios (Table 1) corresponding to different mutational models, different strengths of selection, and different numbers of loci.

We fit the simulated data using model 1 and used the KLD criterion ($p = 0.05$) to identify outliers. A summary of the results is shown in Table 1. Several important features are apparent. First, neutral loci are rarely misclassified as being subject to selection. Only in one set of simulations was the false positive rate higher than 5%. Second, under conditions of the simulation loci subject to balancing selection are rarely detected. Only when the selective advantage of heterozygotes is very strong ($s = 0.2$) and the marker itself is subject to selection do we detect stabilizing selection in more than 30% of cases. Third, loci subject to divergent directional selection are quite readily detected when the selection coefficient is moderate to strong ($s = 0.05$), regardless of whether selection acts directly on the marker or on a tightly linked locus.

[Table 1 about here.]

Thus, if allelic variation at most loci in a sample is selectively neutral and if mutation rates at those loci are the same, loci we designate as statistical outliers correspond to a large fraction of loci that are subject to divergent selection pressures. The lack of power to detect balancing (stabilizing) selection is not surprising. Given our simulation conditions F_{ST} at neutral loci is expected to be about 0.1. Detecting balancing selection would require us to detect loci at which $F_{ST} < 0.1$, which is very difficult given that F_{ST} is bounded below by 0. Detecting divergent selection on the other hand requires detection of loci at which $F_{ST} > 0.1$. Moreover, a reviewer pointed out that the stationary distribution of allele frequencies at such loci depends on Ns, where N is the effective size of local populations and s is the selection coefficient (Wright (1931), see also Holsinger (1999)). Thus, in a situation where local populations consist of 2500 individuals rather than 250, our approach may detect

a large fraction of loci subject to divergent selection even when the selection coefficient is as small as 0.01.

6. APPLICATION AND MODEL COMPARISON

We apply the proposed models to two subsets of SNP data on human chromosome 7: (1) low-resolution data including 3040 loci separated by approximately 53k base pairs and (2) high-resolution data including 3002 loci separated by approximately 860 base pairs (see Section 2 for details). The high-resolution data are centered around SNP $rs13239338$, which has the largest KLD measure identified in the low-resolution analysis.

The proximity matrix is calculated from genetic map positions, as described earlier. Based on the three criteria introduced in section 3, we adopt the following proximity functions for low and high resolution data:

$$c(d_{ij}) = \begin{cases} 0.5/3040 + 0.0125\exp(-|d_{ij}|) & \text{low resolution data} \\ 0.5/3002 + 0.02\exp(-1000 * |d_{ij}|) & \text{high resolution data,} \end{cases}$$

where d_{ij} is the distance (in centimorgans) between loci i and j.

We fit the models using MCMC. Except for a few parameters that can be sampled directly from conditional distributions, most parameters are sampled using the Metropolis-Hastings (M-H) updates. Examination of the trace and standard convergence diagnostics (Geweke 1992) suggest that convergence has been achieved.

We use two criteria to compare models: the Deviance Information Criterion (DIC) and the Conditional Predictive Ordinate (CPO) based log of the Pseudomarginal likelihood (LPML). DIC assesses models on the marginal space and is defined as

$$DIC = \bar{D} + p_D,$$

where D is the Bayesian deviance, $D = -2\log(p(y|\theta)) + 2\log(f(y))$ and $\bar{D}$ is the posterior mean of D. p_D is a penalty term: $p_D = \bar{D} - D(\bar{\theta})$, where $D(\bar{\theta})$ is the Bayesian deviance measured at posterior mean of parameter θ.

The parameters we are interested in, θ_i and ϵ_i, are at the hyperparameter level. Thus, it is more appropriate to assess the model based on θ_i and ϵ_i than based on p_{ik}. In other words, θ_i and ϵ_i are the parameters of focus in the sense of Spiegelhalter, Best, Carlin and van der Linde (2002), and the p_{ik} can be considered nuisance parameters. In light of this, we integrate the p_{ik} out and calculate DIC based on θ_i and ψ_i, i.e.,

$$\begin{aligned}&\int f(\mathbf{x}|\boldsymbol{p})\pi(\boldsymbol{p}|\boldsymbol{\theta},\boldsymbol{\psi})d\boldsymbol{p}\\&=\prod_{i=1}^{I}\prod_{k=1}^{K}\binom{N_{ik}}{x_{ik}}\frac{1}{B(\frac{1-\theta_i}{\theta_i}\psi_i,\frac{1-\theta_i}{\theta_i}(1-\psi_i))}B(x_{ik}+\frac{1-\theta_i}{\theta_i}\psi_i,N_{ik}-x_{ik}+\frac{1-\theta_i}{\theta_i}(1-\psi_i)).\end{aligned} \tag{13}$$

CPO and LPML are model evaluation criteria based on the predictive space (Gelfand and Dey 1994; Gelfand, Dey and Chang 1992; Geisser 1993; Dey, Chen and Chang 1997). The CPO for x_{ik}, the allele count at locus i in population k, is defined as

$$\text{CPO}_{ik}=f(x_{ik}|\mathbf{D}_{(-ik)})=\int f(x_{ik}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{D}_{(-ik)})d\boldsymbol{\theta},$$

where $\mathbf{D}_{(-ik)}$ denotes the data with observation x_{ik} deleted and $\pi(\boldsymbol{\theta}|\mathbf{D}_{(-ik)})$ is the posterior density of the model parameter $\boldsymbol{\theta}$ based on the data $\mathbf{D}_{(-ik)}$. LPML is the summation of the logarithm of the CPOs,

$$\text{LPML}=\sum_{i=1}^{I}\sum_{k=1}^{K}\log(\text{CPO}_{ik}).$$

CPO can be calculated using Monte Carlo approximation directly from the MCMC output ,

$$\text{CPO}_{ik}=\left(\frac{1}{B}\sum_{b=1}^{B}\frac{1}{f(x_{ik}|\boldsymbol{\theta}^{(b)})}\right)^{-1},$$

where $\{\boldsymbol{\theta}^{(b)},b=1,...,B\}$ is the MCMC sample from $\pi(\boldsymbol{\theta}|\mathbf{D})$ and $\mathbf{D}$ is the complete data. As with DIC, the CPO calculation can be based either on p_{ik} or on $(\theta_i,\theta_k,$ and $\psi_i)$. Again, we want to predict the allele counts x_{ik} given the parameter of interest $(\theta_i,\theta_k,\psi_i|\mathbf{D}_{(-ik)})$. The p_{ik} should be considered as random effects rather than model parameters. Therefore, we integrate the p_{ik} out and use equation (13) to calculate CPO and LPML.

Table 2 summarizes DIC and LPML results. For low-resolution SNP data, Model 2 has the smallest DIC thus is preferred to the alternative models. The ordering of models according to the LPML criterion is identical. Both CAR models are inferior to the non-spatial alternatives. Thus, spatial effects are not detectable at low resolution, but population specific effects are important. The lack of spatial effect may not be too surprising in these data, because the average distance between adjacent markers is more than 52kb. Because of these results, outlier detection in the low-resolution data is based on Model 2.

With the high-resolution data, the CAR models outperform models that fail to account for the statistical association expected between loci that are in close proximity. Once the effects of spatial proximity have been accounted for, however, we find no detectable effect associated with population. The lack of population specific effect in these data may not be surprising, because the high-resolution data cover only about 2% of chromosome 7 and are centered on a marker already known to exhibit much more among-population differentiation than the genomic average. Thus, we use Model 1 with a CAR prior to detect outliers in the high-resolution data.

[Table 2 about here.]

To identify outliers in the HapMap data we chose a critical KLD value of 1.614 ($p = 0.01$). Using this criterion we identified 17 loci as outliers (Figure 2). In every case, the posterior distribution of θ_i is substantially shifted to the right, indicating that all of these loci mark regions of the genome showing substantially greater differentiation among populations than the average locus in our sample. Ten of the 17 loci we identify as outliers are located either within or close by a known gene or open-reading frame. The relationship between known genes and loci we identify as outliers is summarized in Table 3.

[Figure 2 about here.]

[Table 3 about here.]

In Table 3, we notice that SNP $rs13239338$ has the highest KLD, even though it is not within a known gene. We use this locus as the center of our high-resolution scan, including 3001 SNP loci around it in the high-resolution data set. Using a critical KLD of 1.614 (p=0.01), we now identify 57 loci showing unusually large amounts of among-population differentiation. Figure 3 shows the genomic location of these θ_i values as well as the locations of known genes in this region. The 57 outliers fall within a smaller number of clusters. Perhaps the most striking cluster is the one involving 16 markers in the vicinity of LEP. A smaller number of markers are clustered around NYD-SP18/CALU and KIAA0828. The remaining markers are spread through a region including GRM8, LOC168850, GCC1 and FSCN3.

We summarize the relationships between known genes and markers identified as outliers in Table 4. It is interesting to observe that SNPs $rs2278815$ and $rs4731426$ are in an intron of LEP leptin, which is the homolog of a gene contributing to obesity in mice. In the Yoruba population, 95% of chromosomes have nucleotide base G at both sites while in other populations the frequency of G is only 22-40%. The protein encoded by this gene is secreted by white adipocytes. In mice, mutations in this gene are associated with severe obesity. The relationship between allelic differences at these SNP loci has also been confirmed in human population association studies (Mammès, Betoulle, Aubert, Herbeth, Siest and Fumeron 2000; Li, Reed, Lee, Xu, Kilker, Sodam and Price 1999; Mammès, Betoulle, Aubert, Giraud, Tuzet, Petiet, Colas Linhart and Fumeron 1998). Our results suggest that allelic differences at other loci involved in fat metabolism must compensate for the among-population allelic differences observed here.

[Figure 3 about here.]

[Table 4 about here.]

7. DISCUSSION

In an earlier analysis of population differentiation using the HapMap data set, Weir et al. (2005) found substantial heterogeneity in locus-specific estimates of F_{ST}. Because their analysis used method-of-moment estimates (Weir and Cockerham 1984; Weir and Hill 2002), they were unable to provide a statistical criterion for recognizing particular loci as outliers, i.e., as exhibiting unusually large or unusually small amounts of differentiation among populations.

In this paper we extend an existing Bayesian framework for analysis of genetic differentiation among populations (Holsinger 1999; Holsinger 2006) to accommodate locus- and population-specific effects on F_{ST}. A novel aspect of our extension is the use of conditional autoregressive models to account for the correlation in patterns of among-population differentiation expected between loci that are in close physical proximity. A model that ignores associations among loci performs well on low-resolution data (one marker every 52kb on average), but including associations among closely linked loci is vital in analyses of the high-resolution data available in the full HapMap data set (one marker every kb on average). We compare estimated locus-specific effects with a hyperdistribution reflecting variation in F_{ST} across all loci in the sample. Strictly speaking, our approach only allows us to identify statistical outliers, but a small simulation study suggests that outliers detected by our approach correspond to loci under selection if most loci in the sample are neutral and share the same or comparable mutation rates.

In our analysis of data derived from the HapMap project, it seems reasonable to conclude that loci we identify as statistical outliers mark regions of the genome that have been subject to divergent selection among the populations included in the sample. For a set of neutral loci, allele frequencies are completely determined by the history of local population sizes they share, the history of migration among local populations they share, and the distribution of mutation rates among loci. If we summarize the amount of genetic differentiation among populations with F_{ST}, then the distribution of F_{ST} across loci will reflect both variation arising from the underlying stochastic evolutionary process and variation arising from differences

among loci in mutation rates. We identified 17 loci in the low-resolution analysis and 57 loci in the high-resolution analysis that are outliers with respect to this hyperdistribution. Such outliers represent loci with levels of among-population differentiation that are substantially larger than would be consistent with the distribution of F_{ST} at the remaining 2900+ loci in our sample, and selection seems more likely than mutation to be responsible for such extreme departures from the genomic average.

We anticipate that our approach to outlier detection is less likely to detect loci that are subject to selection than approaches that directly model the demographic history of populations. Storz et al. (2004), for example, use a coalescent approach to estimate demographic parameters and construct expectations based on those parameter estimates. As discussed by Nielsen (2001; 2005) tests like these depend on strong assumptions about demography. We suspect that making such parametric assumptions make the tests based on coalescent approaches more sensitive to departures from neutrality. In our approach the vagaries of demographic history are shared by all autosomal loci, and the hyperdistribution describing F_{ST} variation among loci encapsulates the uncertainty associated with the drift process. Thus, our method will be robust to a variety of demographic scenarios, although it is likely to be less powerful than methods designed to take those scenarios into account.

Human population geneticists have made a wealth of data available in recent years. The HapMap data set, of which we have analyzed only a small portion here, includes samples only from four geographically distinct populations, but the data from these populations is available at high genomic resolution, roughly every 1kb over the entire human genome. In contrast, the HGDP-CEPH microsatellite data set (Cann, de Toma, Cazes, Legrand, Morel, Piouffre, Bodmer, Bodmer, Bonne-Tamir, Cambon-Thomsen, Chen, Chu, Carcassi, Contu, Du, Excoffier, Friedlaender, Groot, Gurwitz, Herrera, Huang, Kidd, Kidd, Langaney, Lin, Mehdi, Parham, Piazza, Pistillo, Qian, Shu, Xu, Zhu, Weber, Greely, Feldman, Thomas, Dausset and Cavalli-Sforza 2002) provides data at low genomic resolution (377 loci or roughly one every 10^7kb), but it includes samples from 52 geographically defined populations in Africa, Eurasia, Oceania, and the Americas. In addition to providing a much wider

geographic sampling of human diversity and thereby increasing the potential that populations have experienced divergent selection pressures, loci in the HGDP-CEPH microsatellite data set harbor many alleles, and the mutational dynamics of microsatellite loci are quite different from those of SNPs. Our future work will include both high resolution analyses of the entire human genome using data derived from the HapMap project and the development of statistical models appropriate for similar analyses of the HGDP-CEPH microsatellite data.

REFERENCES

Akey, J. M., Zhang, G., Zhang, K., Jin, L., and Shriver, M. D. (2002), "Interrogating a High-Density SNP Map for Signatures of Natural Selection," *Genome Res.*, 12(12), 1805–1814.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, FL: Chapman & Hall/CRC.

Beaumont, M. A., and Balding, D. J. (2004), "Identifying adaptive genetic divergence among populations from genome scans," *Molecular Ecology*, 13(4), 969–980.

Cann, H. M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W. F., Bonne-Tamir, B., Cambon-Thomsen, A., Chen, Z., Chu, J., Carcassi, C., Contu, L., Du, R., Excoffier, L., Friedlaender, J. S., Groot, H., Gurwitz, D., Herrera, R. J., Huang, X., Kidd, J., Kidd, K. K., Langaney, A., Lin, A. A., Mehdi, S. Q., Parham, P., Piazza, A., Pistillo, M. P., Qian, Y., Shu, Q., Xu, J., Zhu, S., Weber, J. L., Greely, H. T., Feldman, M. W., Thomas, G., Dausset, J., and Cavalli-Sforza, L. L. (2002), "A Human Genome Diversity Cell Line Panel," *Science*, 296(5566), 261b–262.

Cavalli-Sforza, L. L. (1966), "Population Structure and Human Evolution," *Royal Society of London Proceedings Series B*, 164, 362–379.

Chakraborty, R., Kimmel, M., Stivers, D. N., Davison, L. J., and Deka, R. (1997), "Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci," *Proceedings of the National Academy of Sciences USA*, 94, 1041–1046.

Consortium, T. I. H. (2005), "A haplotype map of the human genome," *Nature*, 437, 1299–1320.

Crow, J. F., and Kimura, M. (1970), *An Introduction to Population Genetics Theory*, Minneapolis, Minn.: Burgess Publishing Company.

Dey, D. K., Chen, M.-H., and Chang, H. (1997), "Bayesian Approach for Nonlinear Random Effects Models," *Biometrics*, 53, No. 4, 1239–1252.

Fu, R., Gelfand, A. E., and Holsinger, K. E. (2003), "Exact moment calculations for genetic models with migration, mutation, and drift," *Theoretical Population Biology*, 63, 231–243.

Geisser, S. (1993), *Predictive Inference: An Introduction*, Boca Raton, FL: Chapman & Hall/CRC.

Gelfand, A. E., and Dey, D. K. (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3), 501–514.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992), "Model determination using predictive distributions with implementation via sampling-based methods," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press, pp. 147–167.

Geweke, J. (1992), "Evaluating the accuracy of sampling-based approaches to calculating posterior moments," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press.

Holsinger, K. E. (1999), "Analysis of genetic diversity in geographically structured populations: a Bayesian perspective," *Hereditas*, 130, 245–255.

Holsinger, K. E. (2006), "Bayesian hierarchical models in geographical genetics," in *Applications of Computational Statistics in the Environmental Sciences*, eds. J. S. Clark, and A. E. Gelfand, New York, NY: Oxford University Press, pp. 25–37.

Lercher, M. J., and Hurst, L. D. (2002), "Human SNP variability and mutation rate are higher in regions of high recombination," *Trends in Genetics*, 18, 337–340.

Lewontin, R. C., and Krakauer, J. (1973), "Distribution of Gene Frequency as a Test of the Theory of the Selective Neutrality of Polymorphisms," *Genetics*, 74(1), 175–195.

Li, C. C. (1955), *Population Genetics*, Chicago, IL: University of Chicago Press.

Li, W. D., Reed, D. R., Lee, J. H., Xu, W., Kilker, R. L., Sodam, B. R., and Price, R. A. (1999), "Sequence variants in the 5' flanking region of the leptin gene are associated with obesity in women," *Annals of Human Genetics*, 63, 227–234.

Malécot, G. (1948), *Les Mathématiques de l'Hérédité*, Paris, France: Masson et Cie.

Mammès, O., Betoulle, D., Aubert, R., Giraud, V., Tuzet, S., Petiet, A., Colas Linhart, N., and Fumeron, F. (1998), "Novel polymorphisms in the 5' region of the LEP gene: association with leptin levels and response to low-calorie diet in human obesity," *Diabetes*, 47, 587–489.

Mammès, O., Betoulle, D., Aubert, R., Herbeth, B., Siest, G., and Fumeron, F. (2000), "Association of the G-2548A polymorphism in the 5' region of the LEP gene with overweight," *Annals of Human Genetics*, 64, 391–394.

Nei, M., and Maruyama, T. (1975), "Lewontin-Krakauer test for neutral genes," *Genetics*, 80(2), 395.

Nielsen, R. (2001), "Statistical tests of selective neutrality in the age of genomics," *Heredity*, 86(6), 641–647.

Nielsen, R. (2005), "Molecular signatures of natural selection," *Annual Review of Genetics*, 39(1), 197–218.

Peng, F., and Dey, D. K. (1995), "Bayesian analysis of outlier problems using divergence measures," *Canadian Journal of Statistics*, 23, 194–213.

Riebler, A., Held, L., and Stephan, W. (2008), "Bayesian variable selection for detecting genomic differences among populations," *Genetics*, (in press).

Robertson, A. (1975), "Gene Frequency Distributions as a Test of Selective Neutrality," *Genetics*, 81(4), 775–785.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002), "Genetic Structure of Human Populations," *Science*, 298(5602), 2381–2385.

Song, S., Dey, D. K., and Holsinger, K. E. (2006), "Differentiation among populations with migration, mutation, and drift: implications for genetic inference," *Evolution*, 60, 1–12.

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2002), "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B*, 64 No.4, 583–639.

Storz, J. F., Payseur, B. A., and Nachman, M. W. (2004), "Genome Scans of DNA Variability in Humans Reveal Evidence for Selective Sweeps Outside of Africa," *Mol Biol Evol*, 21(9), 1800–1811.

Weber, J. L., and Wong, C. (1993), "Mutation in short tandem repeat polymorphisms," *Human Molecular Genetics*, 2, 1123–1128.

Weir, B. S., Cardon, L. R., Anderson, A. D., Nielsen, D. M., and Hill, W. G. (2005), "Measures of human population structure show heterogeneity among genomic regions," *Genome Res*, 15(11), 1468–76.

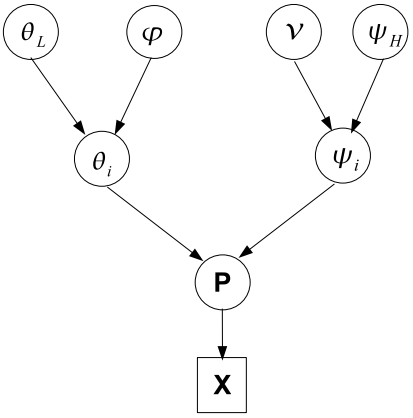
Weir, B. S., and Cockerham, C. C. (1984), "Estimating F-statistics for the analysis of population structure," *Evolution*, 38, 1358–1370.

Weir, B. S., and Hill, W. G. (2002), "Estimating F-statistics," *Annual Reviews of Genetics*, 36, 721–750.

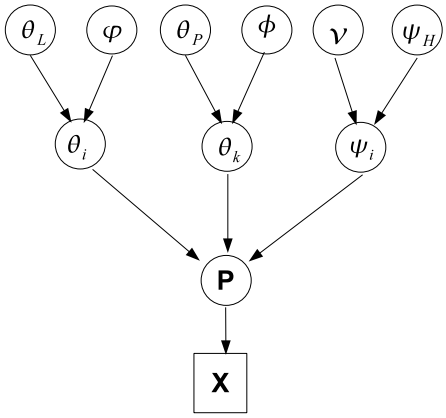
Wright, S. (1931), "Evolution in Mendelian populations," *Genetics*, 16, 97–159.

Wright, S. (1951), "The genetical structure of populations," *Annals of Eugenics*, 15, 323–354.

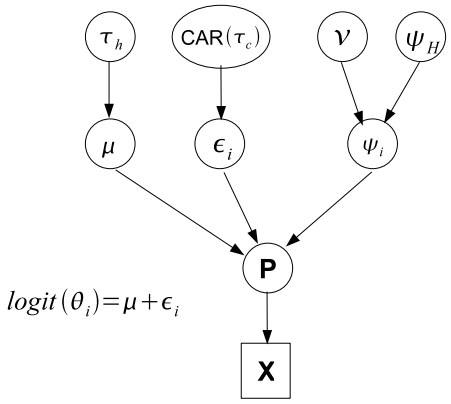
List of Figures



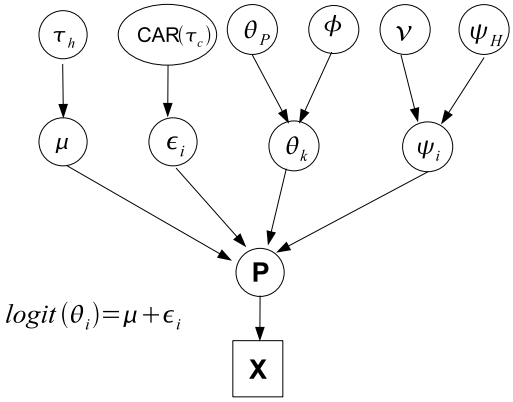
Model 1



Model 2



Model 1 with CAR prior



Model 2 with CAR prior

Figure 1: DAG plot of the models

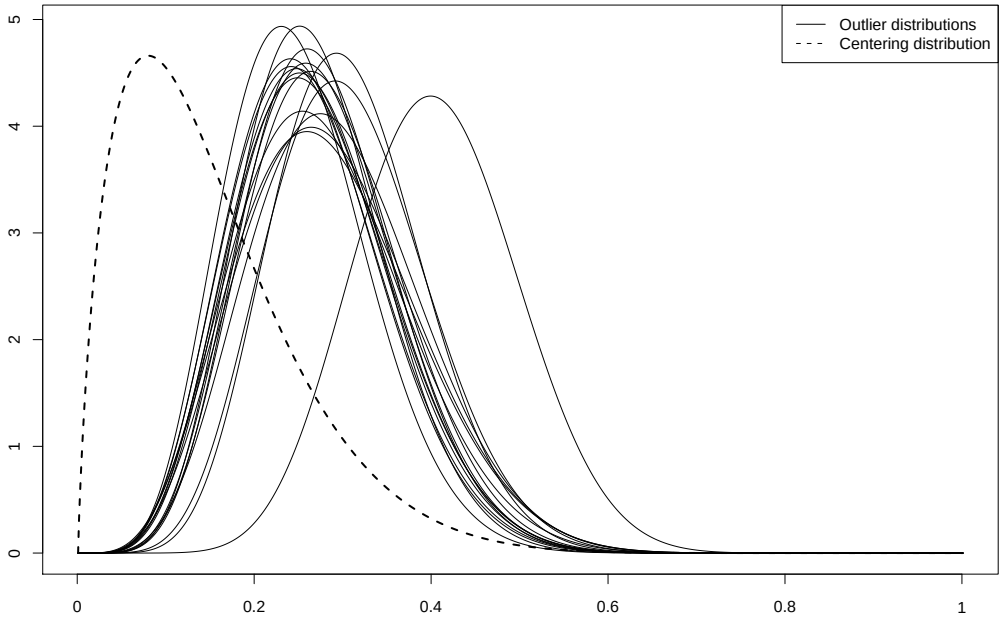


Figure 2: Densities of posterior θ_i for low resolution scan

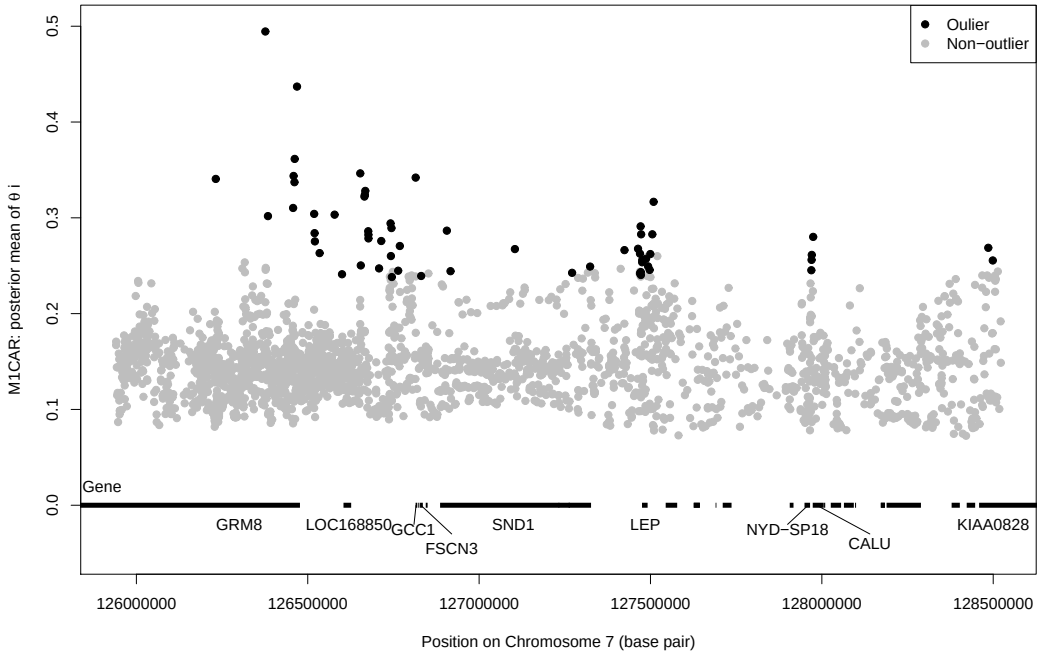


Figure 3: High resolution scan outliers (M1CAR)

List of Tables

Table 1: Simulation results

Selection model	Selection coefficient	Number of loci			Classification*		
		Directional	Balancing	Neutral	Direc.	Bal.	Neut.
2-locus	0.02	80	20	900	**4%**	0%	0%
					0%	**0%**	0%
					96%	100%	**100%**
2-locus	0.05	80	20	900	**68%**	0%	0%
					0%	**0%**	0%
					32%	100%	**100%**
2-locus	0.1	80	20	900	**90%**	0%	1%
					0%	**0%**	1%
					10%	100%	**98%**
2-locus	0.02	40	10	450	**10%**	0%	1%
					0%	**0%**	0%
					90%	100%	**99%**
2-locus	0.05	40	10	450	**68%**	0%	0%
					0%	**0%**	0%
					32%	100%	**100%**
2-locus	0.1	40	10	450	**88%**	0%	2%
					0%	**0%**	0%
					12%	100%	**98%**
2-locus	0.2	40	10	450	**100%**	0%	3%
					0%	**30%**	4%
					0%	70%	**93%**
Marker selected	0.02	40	10	450	**0%**	0%	0%
					0%	**0%**	0%
					100%	100%	**100%**
Marker selected	0.05	40	10	450	**63%**	0%	1%
					0%	**0%**	0%
					27%	100%	**99%**
Marker selected	0.1	40	10	450	**90%**	0%	1%
					0%	**20%**	0%
					10%	80%	**99%**
Marker selected	0.2	40	10	450	**88%**	0%	0%
					0%	**50%**	5%
					12%	50%	**95%**

*Column is the true scenario and row is the classified scenario. Bold numbers indicate correct classification (using critical KLD=0.830, $p = 0.05$).

Table 2: Model comparison

	Low-resolution data				High-resolution data			
Model	$\bar{D}$	p_D	DIC	LPML	$\bar{D}$	p_D	DIC	LPML
M1	81225	2809	84034	-42695	71205	2579	73784	-37423
M2	**80417**	**2776**	**83192**	**-42356**	71291	2541	73833	-37436
M1CAR	81145	3141	84285	-42690	**70968**	**2813**	**73780**	**-37389**
M2CAR	80534	3166	83700	-42368	71261	2557	73818	-37398

Table 3: Outliers detected in low-resolution data

SNP ID	Position	KLD (M2)	Candidate loci
rs7787411	14746124	2.35	diacylglycerol kinase
rs10263500	30557791	2.20	corticotropin releasing hormone receptor 2, indolethlyamin N-methyltransferase
rs11771444	30797413	1.76	growth hormone releasing receptor
rs12535578	54928795	3.51	epidermal growth factor receptor
rs4521648	70374286	1.73	UDP-GalNAc:polypeptide
rs2722963	82702679	1.70	SEMA3E: semaphorin 3E
rs1990040	85957725	1.83	glutamate receptor
rs17161695	98609357	2.29	actin-related protein 2/3 complex subunit 1A, PDGFA associated protein 1
rs11976018	98767088	1.87	zinc finger protein 95 homolog (mouse)
rs1476471	108525466	1.65	n.a.[1]
rs43083	111841138	1.72	n.a.[1]
rs12531918	111885313	2.40	n.a.[1]
rs2894673	112653596	1.79	n.a.[1]
rs6466707	118915526	1.92	n.a.[1]
rs13239338	126578324	7.21	n.a.[1]
rs2671095	131284016	3.07	n.a.[1]
rs4716934	155227668	2.17	Homo sapiens sonic hedgehog homolog (Drosophila)

[1]No known genes in vicinity of this SNP.

Table 4: High-resolution data outliers and known genes

Gene & Location	SNP loci
GRM8: glutamate receptor, (125672607,126477260)	rs7796270, rs7786541, rs17865314, rs6960871, rs4532535, rs2106149, rs2237808, rs10226369
LOC168850: hypothetical protein (126604306:126626717)	rs951809
GCC1: Golgi coiled-coil protein 1 (126819604,126814633)	rs989100
FSCN3: fascin3 (126827639:126835793)	rs806214
SND1:staphylococcal nuclease domain containing 1 (126886152:127326608)	rs7793281, rs712707, rs12672945, rs6969233, rs322821
LEP:leptin precursor (obesity homolog, mouse) (127475281:127491631)	rs2021808, rs4731423, rs4731424, rs1349419, rs13245201, rs10487506, rs7799039, rs2278815, rs4731426, rs2071045, rs2060715, rs4731429, rs10954175, rs12537998, rs1466145, rs4728090,
NYD-SP18: testes development-related (127949393 :127965611) CALU: calumenin recursor (127973386:128005477)	rs17164371, rs2402934, rs7780294, rs2060717
KIAA0828: adenosylhomocysteinase 3 (128458814: 128664001)	rs721691, rs4731568
Loci with no known gene within 5k bps (around LOC168850)	rs1419391, rs975308, rs916598, rs12536774, rs13239338, rs9640842, rs2106177, rs10487482, rs4731365, rs12666432, rs12673058, rs10954158, rs1419410, rs12671806, rs12668127, rs11768389, rs1592365, rs11984364, rs17150996